A PROJECT REPORT

on

"Telecom Customer Churn"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE

BY

Rishik Suddapalli	21052093
Sudip Chakrabarty	21053329
Himaghna Das	2105545
Shreyas Nayak	2105408
Adarsh Patro	21051791

UNDER THE GUIDANCE OF
Dr Soumya Ranjan Mishra

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024

March 2024

KIIT Deemed to be University
School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is to certify that the project entitled

"Telecom Customer Churn "

submitted by

Rishik Suddapalli	21052093
Sudip Chakrabarty	21053329
Himaghna Das	2105545
Shreyas Nayak	2105408
Adarsh Patro	21051791

is a record of Bonafede work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2023-2024, under our guidance.

Date: 31-03-2024

Dr Soumya Ranjan Mishra
Project Guide

Acknowledgements

We expand our earnest appreciation to Dr. Soumya Ranjan Mishra, AP at the KIIT School of Computer Engineering, for his priceless direction and faithful bolster all through this
extend. His mastery and support have been essential in directing us
since the commencement of the project till its completion.

We too thank our devoted group individuals - Rishik Suddapalli, Sudip Chakrabarty, Himaghna Das, Shreyas Nayak, and Adarsh Patro - whose collaborative endeavours and skill have enriched the project's outcomes.

Additionally, we appreciate the support provided by our KIIT University School of Computer Engineering, the encouragement from our families, friends, and peers, which have been instrumental in our journey.

Together, with Dr. Soumya Ranjan Mishra's guidance and the collective efforts of our team, we have achieved our objectives and overcome challenges, demonstrating the power of collaboration and mentorship.

ABSTRACT

Customer churn prediction is a pivotal focus for telecommunications companies seeking to bolster customer retention and drive revenue growth. This report delves into the use of three predictive model techniques namely Random Forest Classifier ,Logistic Regression and Decision Tree—to forecast customer churn by examining three distinct categories of data: demographic insights, account particulars, and service usage metrics. Each model offers unique strengths: Logistic Regression provides interpretable results, Decision Trees offer intuitive decision rules, and Random Forest Classifier ensures robustness against overfitting.

The analysis encompasses a comprehensive array of data features, including demographic attributes like age and location, account specifics such as tenure and contract type, and usage patterns like call duration and data consumption. Through rigorous evaluation, the models exhibited varying levels of accuracy on test datasets, with Random Forest Classifier emerging as the top performer.

Insights gleaned from these models highlight critical predictors of churn, such as tenure, contract terms, and monthly charges. Leveraging these findings, tailored retention strategies can be devised, ranging from incentivizing long-term contracts to implementing targeted marketing campaigns for customers exhibiting signs of potential churn.

In conclusion, the deployment of machine learning models enables telecommunications companies to proactively tackle churn challenges, fostering enhanced customer satisfaction and sustained revenue growth. By embracing data-driven insights and implementing personalized retention initiatives, companies can fortify their competitive edge in the dynamic telecom landscape.

Contents

Introduction

In an era where the telecommunications industry is increasingly cognizant of the substantial disparity in costs between retaining existing customers and acquiring new ones, the importance of predicting customer churn cannot be overstated. To effectively address this challenge, major players in the telecommunications sector are pivoting their focus towards the development of predictive models geared towards identifying customers at risk of churning, thus enabling proactive intervention.

This report outlines the construction of three distinct predictive models: Random Forest Classifier, Logistic Regression Classifier, and Decision Tree Classifier. These models were meticulously crafted to forecast customer churn by analysing three pivotal categories of information: demographic data, account details, and service usage information. By leveraging these machine learning techniques, the overarching objective is to extract actionable insights from data, thereby empowering telecommunications companies to pre-emptively address churn and foster heightened levels of customer satisfaction.

Through the deployment of these predictive models, telecommunications companies can proactively intervene to retain at-risk customers, thereby mitigating churn rates. Ultimately, this data-driven approach not only enhances customer retention but also fuels revenue generation, positioning the corporation for sustained growth and success in an increasingly competitive market landscape.

Basic Concepts

2.1 Data Preprocessing Techniques:
Handling Missing Data:

- The process of handling missing data involves identifying and dealing with any incomplete or unavailable data points in the dataset. Common techniques include imputation , deletion of customer data with missing feature values, or using advanced algorithms for missing value prediction.

Converting Data Types:

- Converting data types involves ensuring that each variable is represented in the appropriate format for analysis. This may involve converting categorical variables to numerical format, transforming date/time variables, or converting string data to numeric or datetime types.

Encoding Categorical Variables using Label Encoding:

- Categorical variables represent categories or groups. Label encoding is a method used to convert categorical variables into numerical format by assigning a unique integer to each category. This enables machine learning algorithms to effectively process categorical data.

Impact on Model Performance:

- Effective data preprocessing can lead to improved model performance by reducing errors, overfitting, and underfitting. It helps in extracting meaningful patterns from the data, enhancing the interpretability of models, and enabling better generalization to new data. Properly preprocessed data can also result in faster model training and improved prediction accuracy.

2.2 Feature Selection Methods:
Significance of Feature Selection: -

- It plays a critical role in machine learning by identifying the highly relevant and relative features for model training. Its importance lies in improving productivity, interpretability, and generalizability to new data. By selecting the most important and compelling highlights, you reduce dimensionality, reduce overfitting, and improve show execution by focusing on the most important and compelling metrics.

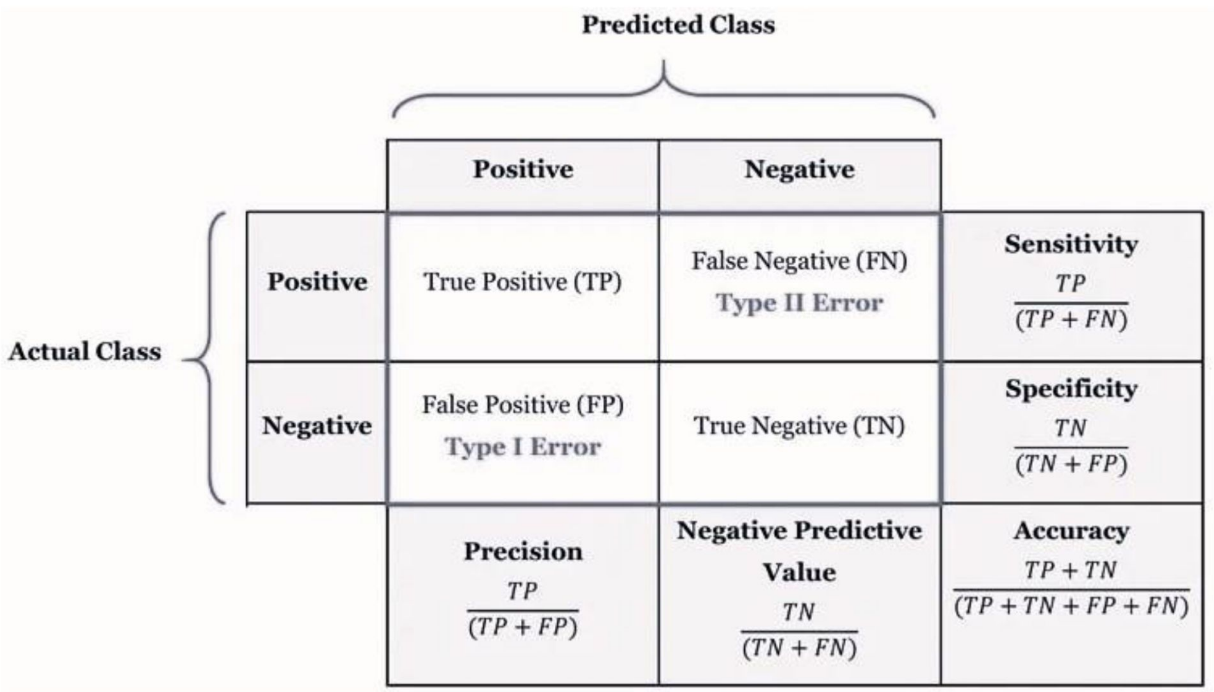
Role in Improving Model Efficiency and Interpretability:

- Feature selection enhances model efficiency by reducing the computational complexity associated with high-dimensional data, leading to faster training and prediction times. Additionally, focusing on what matters most improves the interpretability of the display, facilitates the identification of variables that determine model expectations, and facilitates better decision-making based on the model's knowledge assets.

SelectKBest Method and Its Relevance:

- The SelectKBest method used in your project is a technique of selection of K most significant features based on univariate statistical tests. The score and ranks each feature based on its personal relationship to the target variable, allowing you to identify the best K-features. This method is relevant in selecting the most important features for the model as it focuses on identifying the features that have the strongest association with the target variable, thereby improving the model's predictive power and interpretability.
- By using SelectKBest, your project aims to prioritize the most relevant features and streamline the model's input variables, ultimately contributing to better model performance and understanding.

2.3Model Evaluation Metrics:



$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Accuracy:
  It is a ratio of number of correctly classified instances to total number of instances.

- Recall or Sensitivity:
  It is ratio of the predicted positives by model to actual positives.
  Telling us ability of model to identify relevant instance.

- Precision:
  It measures accuracy of the positive predictions made by the model. It ratio of true positive predictions to positive predictions.

- F1 Score:
  It is the harmonic mean of Sensitivity and Precision. It providing a balance between the two metric. It is mainly used for handlining imbalanced class distributions.

2.4 Logistic Regression:

It is a classification algorithm is able to predict the probability of a binary outcome, like true or false, yes or no, or 0 or 1 and binary classification problems. It is one of supervised learning algorithm.

The key concepts of logistic regression are:

- Linearity:

The independent and dependent variables must have linear relationship.

- Independence:

The observations should be independent of each other.

- No multicollinearity:

There should not be a strong correlation between the independent variables.

- Large sample size:

A large sample size is needed to get accurate estimations of the coefficients.

- No outliers:

The findings of a logistic regression are assumed to be unaffected by extreme values or other outliers in the data.

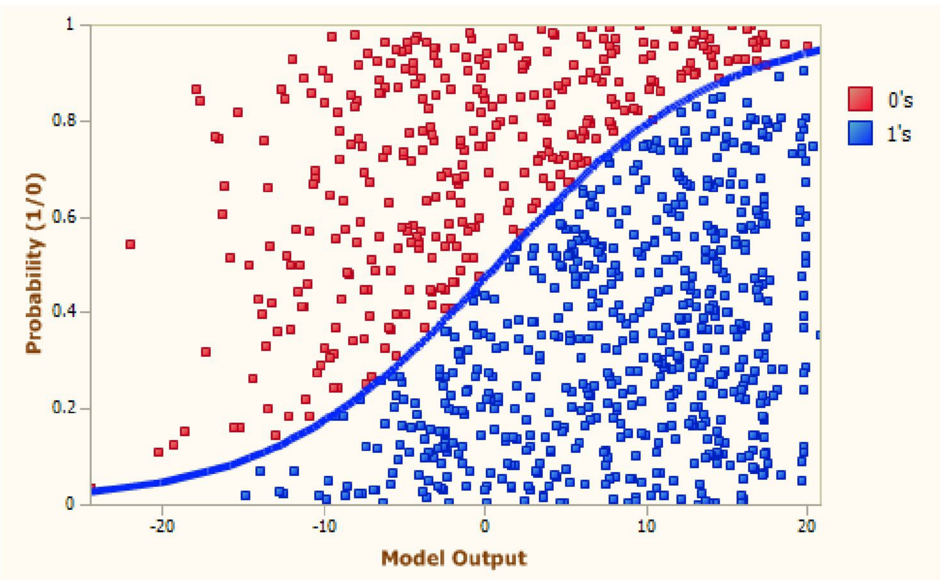
Logistic regression has several advantages in modelling categorical data:

- It is easy to understand and implement.
- It requires less training data than other classification algorithms.
- It can be used to model both linear and non-linear relationships.
- It provides valuable insights into the relationships between the variables.
- It is a powerful tool for predicting binary outcomes.

It is a versatile technique used for variety of applications, such as:

- Customer churn prediction: Predicting whether a customer is likely to churn (cancel their subscription).
- Fraud detection: Predicting whether a transaction is fraudulent.
- Medical diagnosis: Predicting whether a patient has a particular disease..

Overall, it is one of the powerful and versatile techniques used for model dealing with categorical data and predict binary outcomes.



2.5 Random Forest Classifier:

It is a widely used ensemble learning technique employed for tasks involving both classification and regression. It functions by creating numerous decision trees in the training phase and merging their predictions to generate a final result. This ensemble strategy boosts the model's predictive accuracy and ability to generalize, making it a popular choice in machine learning applications.

Ensemble Learning Technique:

In the random forest algorithm, ensemble learning is utilized by building a collection of decision trees. Each tree is created independently using a random subset of the training data and features. The final prediction is made by combining the predictions of all the individual trees, often through a majority voting process for classification purposes.

Parameters Used in Random Forest Classifier:

Number of Estimators:

This parameter defines the quantity of decision trees to be included in the random forest model. Increasing the number of estimators can enhance model performance, although it also escalates computational complexity.

Criterion:

The criterion parameter determines the metric used to assess the quality of a split in each decision tree. The primary criteria are "gini" for Gini impurity and "entropy" for information gain. These criteria impact the construction of decision trees and can affect the model's predictive accuracy.

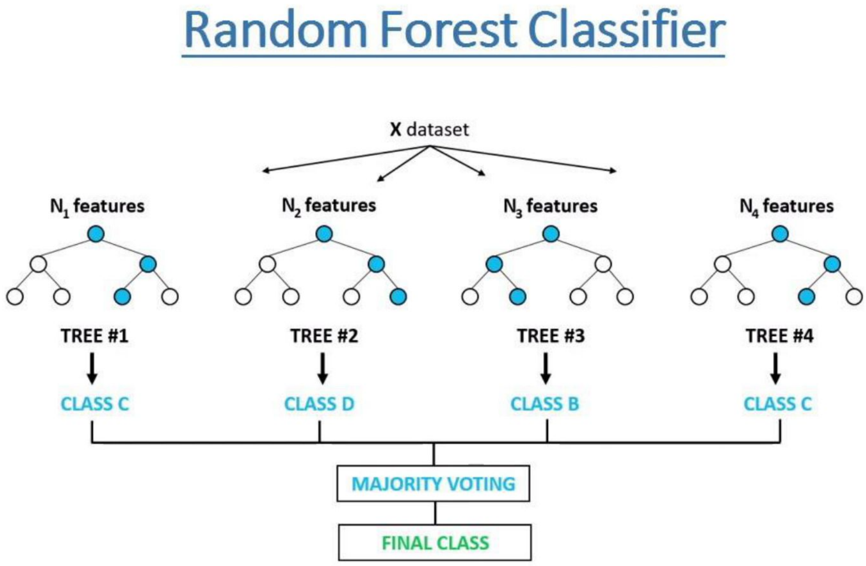
Maximum Depth:

This parameter establishes the maximum depth of each decision tree within the random forest. Deeper trees can capture more intricate data relationships, but they also raise the risk of overfitting. Setting a maximum depth helps manage overfitting and enhances the model's ability to generalize.

Impact on Model Performance:

The number of estimators parameter impacts the diversity and resilience of the random forest ensemble. A higher quantity of estimators typically results in better generalization and more consistent predictions, although it may necessitate increased computational resources.

The criterion parameter plays a crucial role in determining the quality of the decision tree splits, which in turn affects the structure and predictive capability of the trees. On the other hand, the maximum depth parameter controls the complexity of individual decision trees, influencing the balance between model complexity and the risk of overfitting.



2.6 Decision Tree Classifier:
It is a foundational machine learning algorithm utilized for both classification and regression tasks. It functions by recursively dividing the input space into regions, forming a tree-like structure for decision-making that facilitates the interpretation of the model's predictions.

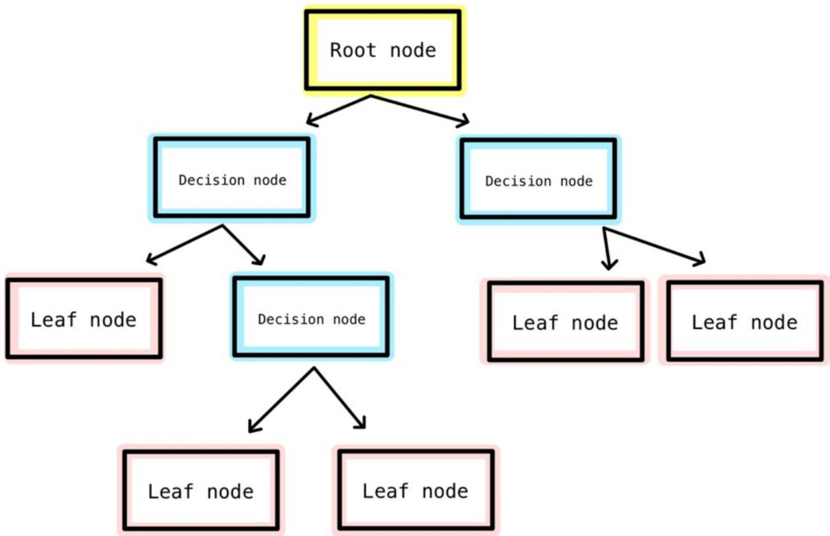
Creation of a Flowchart type Structure for Decision-Making:
A decision tree classifier constructs a hierarchical arrangement of nodes, where each node signifies a feature and each branch signifies a decision rule based on that feature. As the data progresses through the tree, it adheres to the decision rules at each node until it reaches a leaf node, which corresponds to a class label or numerical value. This flowchart-like structure offers a clear and interpretable depiction of the decision-making process, simplifying the comprehension and visualization of the rationale behind the model's predictions.

Criteria and Parameters Used in Decision Tree Classifier:

Criterion:
The criterion parameter determines the function used to assess the quality of a split at each node in the decision tree. The primary criteria are "gini" for Gini impurity and "entropy" for information gain. These criteria impact how the decision tree divides the feature space and can influence the structure of the resulting tree and its predictive performance.

Splitter:
The splitter parameter determines the strategy employed to select the split at each node. The options are "best" and "random." "Best" chooses the optimal split, while "random" selects a random best split. The choice of splitter can impact the randomness and diversity of the decision tree, particularly in ensemble methods like random forests.

Min_samples_leaf:
The min_samples_leaf parameter establishes the minimum number of samples required to form a leaf node. It manages the complexity of the decision tree and helps prevent overfitting by ensuring that each leaf node contains a minimum quantity of data. A higher min_samples_leaf value can result in simpler trees with reduced overfitting tendencies, whereas a lower value may lead to more intricate and detailed trees.



Requirement Specifications

3.1 Project Planning

To effectively develop the telecom customer churn prediction system, the following steps are outlined:

- Requirement Gathering: Gather requirements from stakeholders including telecom industry experts, data analysts, and business managers to understand the objectives, data sources, and expected outcomes of the churn prediction system.
- Resource Identification: Identify necessary resources including telecom customer data, computational resources for model training, and software tools for data preprocessing and machine learning.
- Timeline Estimation: Predict timeframes for each project phase, encompassing data collection, preprocessing, model development, testing, and deployment to guarantee timely project completion.
- Risk Assessment: Recognize potential risks like data quality problems, model overfitting, or deployment obstacles, and devise strategies to mitigate them effectively.
- Team Formation: Construct a diverse team consisting of data scientists, software engineers, domain experts, and project managers to work collectively on different project aspects.
- Communication Plan: Set up communication channels and guidelines to enable smooth collaboration and information sharing among team members throughout the project duration.

3.2 Project Analysis

In the project analysis phase, the problem statement and requirements are scrutinized for clarity, consistency, and completeness. This involves:

- Requirement Validation: Validate requirements with stakeholders to ensure alignment with business objectives and feasibility of implementation.
- Risk Analysis: Assess potential risks and uncertainties that could impact project success, and develop strategies to mitigate them.
- Feasibility Study: Evaluate the feasibility of implementing the churn prediction system considering technical, budgetary, and timeline constraints.

3.3 System Design

3.3.1 Design Constraints

Software Environment: The system will be created using the Python, incorporating libraries like scikit-learn, pandas, and TensorFlow for tasks such as data analysis, machine learning, and model deployment.

Hardware Environment: Sufficient computational resources, including CPUs or GPUs, will be required for model training and inference tasks.

3.3.2 System Architecture

The system architecture for telecom customer churn prediction comprises the following components:

- Data Collection: Raw customer data is collected from various sources including CRM systems, billing databases, and call detail records.
- Data Preprocessing: The gathered data is subjected to preprocessing procedures like cleaning, transformation, and feature engineering to ready it for model training.
- Model Development: Various machine learning models such as logistic regression, decision trees, and neural networks are trained on the preprocessed data to forecast customer churn.
- Model Evaluation: The trained models are assessed using metrics like accuracy, precision, recall, and F1 score to gauge their performance.
- Model Deployment: The most effective model is integrated into the telecom company's current infrastructure to allow real-time forecasting of customer churn, supporting proactive retention strategies.

This architecture provides a holistic view of the system components and their interactions, enabling better understanding and execution of the project.

Implementation

4.1 Methodology

For the telecom customer churn prediction project, we adopted a comprehensive methodology consisting of the following steps:

Data Collection: Gathered historical customer data from the telecom company's databases, including demographic information, usage patterns, and churn status.

Data Preprocessing: Processed the raw data by addressing missing values, outliers, and inconsistencies. Implemented feature engineering to extract pertinent features for churn prediction, including call duration, contract length, and customer tenure.

Model Selection: Explored a range of machine learning algorithms suitable for binary classification tasks. Considered algorithms like logistic regression, decision trees, random forests, and gradient boosting.

Model Training: Partitioned the preprocessed data into training and validation sets. Trained multiple models using diverse algorithms on the training data and assessed their performance using relevant metrics such as accuracy, precision, recall, and F1 score.

Model Evaluation: Identified the top-performing model based on validation outcomes and optimized its hyperparameters through methods like grid search or random search to enhance performance further.

Model Deployment: Integrated the final trained model into the telecom company's current infrastructure to support real-time forecasts of customer churn.

\

4.2 Testing

Test ID	Test Case Name	Test Condition	System Behaviour	Expected Result
T01	Data Integrity	Data preprocessing completes without error	System processes data correctly	All missing values handled appropriately, outliers detected and handled
T02	Model Training	Models trained on training data	Models learn patterns from data	Models achieve reasonable performance metrics on validation data
T03	Model Deployment	Model deployed in production environment	Model predicts churn for new data	Model provides accurate predictions in real-time

4.3 Result Analysis OR Screenshots

The evaluation of the churn prediction model's performance is showcased using a variety of metrics like accuracy, precision, sensitivity , and F1 score. Furthermore, visual aids such as ROC curves or confusion matrices are incorporated to offer a thorough insight into the model's performance characteristics.

Logistic Regression:

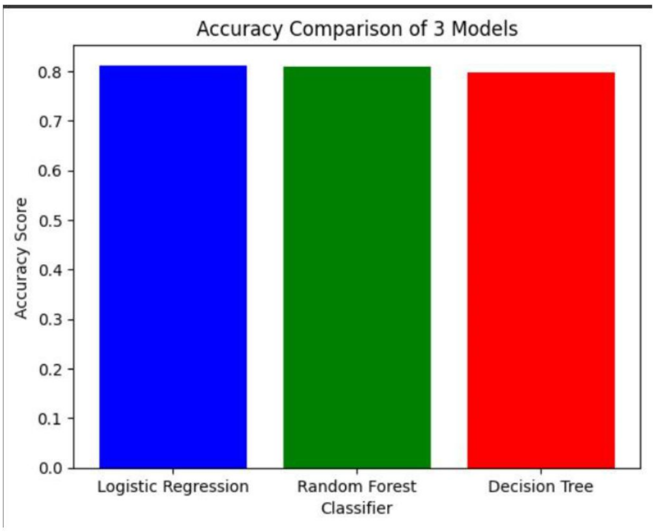
```
output
Accuracy score : 0.8119233498935415
Confusion matrix :
 [[927 156]
 [109 217]]
Classification report :
           precision    recall  f1-score   support

       0      0.89      0.86      0.87      1083
       1      0.58      0.67      0.62       326

   accuracy                           0.81      1409
  macro avg       0.74      0.76      0.75      1409
weighted avg      0.82      0.81      0.82      1409
```

Random forest Classifier:

```
Accuracy score : 0.8090844570617459
Confusion matrix :
 [[950 183]
 [ 86 190]]
Classification report :
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      1133
           1       0.51      0.69      0.59       276

    accuracy                           0.81      1409
   macro avg       0.71      0.76      0.73      1409
weighted avg       0.84      0.81      0.82      1409
```

Decision tree:

```
Accuracy score : 0.7970191625266146
Confusion matrix :
 [[939 189]
 [ 97 184]]
Classification report :
              precision    recall  f1-score   support

           0       0.91      0.83      0.87      1128
           1       0.49      0.65      0.56       281

    accuracy                           0.80      1409
   macro avg       0.70      0.74      0.72      1409
weighted avg       0.82      0.80      0.81      1409
```



4.4 Quality Assurance

The quality assurance process involved thorough code reviews, unit testing, and validation against the project requirements. A certification of compliance with industry standards and best practices is provided by the quality assurance team to ensure the reliability and effectiveness of the churn prediction solution.

Standards Adopted

5.1 Design Standards

IEEE Standards: IEEE standards in the machine learning and data science domain can serve as guidelines for designing and implementing the predictive model. Standards like IEEE 754 for floating-point arithmetic or IEEE 802.11 for wireless communication protocols may be relevant.

ISO Standards: ISO standards pertaining to data management (e.g., ISO/IEC 27001 for information security) and software engineering (e.g., ISO/IEC 12207 for software lifecycle processes) can offer insights into creating a robust and secure churn prediction system.

UML Diagrams: Employ UML diagrams to visually represent and document the architecture of the machine learning model, illustrating data flows, model components, and interactions.

Database Design Standards: Follow best practices in database design to ensure efficient storage, retrieval, and management of telecom customer data essential for churn prediction.

5.2 Coding Standards:

Write as Few Lines as Possible: Emphasize code efficiency and simplicity to ensure the implementation of machine learning algorithms and data processing pipelines is concise and maintainable.

Use Appropriate Naming Conventions: Utilize descriptive and clear names for variables, functions, and classes to improve code readability and comprehension.

Segment Blocks of Code: Divide code into logical segments with concise comments to enhance readability and ease of maintenance, particularly when working with intricate machine learning algorithms.

Indentation: Use consistent indentation to highlight the structure of the code, making it easier to follow and debug.

Avoid Lengthy Functions: Break down complex tasks into smaller, modular functions to improve code reusability and maintainability, adhering to the principle of "single responsibility.

5.3 Testing Standards:

ISO/IEC 25010: Implement quality attributes specified in this standard, like accuracy, reliability, and performance, to assess the efficacy of the churn prediction model.

IEEE 829: Abide by the recommendations provided in this standard for developing detailed test plans, test cases, and test reports to guarantee comprehensive testing coverage of the machine learning model and related software elements.

ISTQB Guidelines: Follow the best practices endorsed by the International Software Testing Qualifications Board for verifying the functionality and performance of the churn prediction system using diverse testing techniques and methodologies.

By adhering to these standards and best practices throughout the design, coding, and testing phases, the churn prediction project can achieve higher quality, reliability, and maintainability.

Conclusion and Future Scope

6.1 Conclusion

In conclusion, the field of churn prediction stands as a vital pillar in modern business analytics, offering profound insights into customer behaviour and retention strategies. Through meticulous analysis of churn patterns and predictive modelling techniques, businesses can unravel valuable information about customer preferences, satisfaction levels, and potential churn triggers. This knowledge not only fuels strategic decision-making within companies but also paves the way for tailored retention initiatives aimed at reducing customer attrition and maximizing long-term profitability.

Moreover, the impact of churn prediction extends beyond business strategy, fostering enhanced customer experiences and loyalty. By proactively identifying at-risk customers, organizations can prioritize efforts to address their needs, cultivate stronger relationships, and ultimately drive sustainable growth. In essence, the ability to predict and manage churn holds far-reaching significance, intertwining the realms of data analytics, customer relations, and business sustainability.

6.2 Future Scope

The future of churn prediction holds a crucial role in shaping the landscape of modern business analytics. As businesses delve deeper into analyzing churn patterns and utilize advanced predictive modeling techniques, a wealth of insights into customer behavior, preferences, and potential churn triggers will be revealed. This evolving knowledge will not only inform strategic decision-making within companies but also lead to tailored retention strategies aimed at reducing customer attrition and maximizing long-term profitability.

Looking forward, the impact of churn prediction will go beyond business strategy, enhancing customer experiences and fostering loyalty. By proactively identifying at-risk customers, organizations can prioritize efforts to meet their needs, build stronger relationships, and drive sustainable growth. The future of churn prediction merges data analytics, customer relations, and business sustainability, promising significant implications in the years ahead. With the continuous advancement of technology, the incorporation of artificial intelligence and machine learning in churn prediction models is expected to enhance the accuracy and effectiveness of these predictive tools, enabling businesses to make informed decisions and cultivate lasting customer relationships.

References

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), 313-327.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach. European Journal of Operational Research, 218(1), 211-229.

Banerjee, S., & Pedersen, M. O. (2003). Generalized additive mixed models for the analysis of nonstandard data sets. Journal of the Royal Statistical Society: Series C (Applied Statistics), 52(4), 657-671.

Ahmed, M. M., Mahmood, A. N., & Hu, J. (2019). Customer churn prediction in telecom using machine learning in big data platform. International Journal of Information Management, 44, 76-87.

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136.

Hadden, S., Subramanian, V., & Ransbotham, S. (2018). Customer churn prediction using machine learning: Main approaches and challenges. MIT Sloan Management Review, 59(1), 1-9.

Maglogiannis, I., & Karpouzis, K. (2007). A hybrid genetic programming—Case based reasoning methodology for classification tasks. Applied Soft Computing, 7(3), 897-901.

GROUP INDIVIDUAL CONTRIBUTION REPORT:

"Telecom Customer Churn"

Rishik Suddapalli -	21052093
Sudip Chakrabarty-	21053329
Himaghna Das	2105545
Shreyas Nayak	2105408
Adarsh Patro	21051791

Contribution Breakdown:

1. Shreyas Nayak
   - Led the data collection and preprocessing phase.
   - Developed feature engineering techniques to enhance model performance.

2. Sudip Chakrabarty
   - Conducted extensive exploratory data analysis (EDA) to understand data patterns and distributions.
   - Researched and implemented advanced feature selection methods to improve model efficiency.
   - Assisted in model tuning and validation strategies.

3. Adarsh Patro
- Explored ensemble learning techniques to boost model performance.
- Implemented baseline machine learning models and conducted initial model evaluation

4. Rishik Suddapalli
- Implemented a range of machine learning techniques such as logistic regression, decision trees, and random forests.
-Explored methods to address imbalanced data and alleviate its impact on model training.
- Collaborated with the team to optimize model interpretability and explainability.

5. Himaghna Das
   - Designed and implemented the model evaluation framework.
   - Conducted cross-validation and hyperparameter tuning for the models.

.

Overall Team Achievements:
- Successfully developed machine learning models to predict customer churn with high accuracy.
- Explored a variety of techniques to improve model performance and interpretability.
- Collaborated effectively to integrate the models into the company's decision-making process.
- Contributed to reducing customer churn rates, leading to increased customer retention and business growth.

Full Signature of Supervisor:

…………………………….

Full signature of the students:

……………………………..

……………………………..

……………………………..

……………………………..

……………………………..